DOCUMENT RESUME

ED 190 639 TM 800 434

AUTHOR Cross, Lawrence H.: And Others
TITLE A New Method for Administering and Scoring Multiple-Choice Tests: Theoretical Considerations and Empirical Results.
PUB DATE Apr 80
NOTE 17p.: Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 8-10, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Answer Sheets: *Guessing (Tests): Higher Education: *Multiple Choice Tests: *Response Style (Tests): *Scoring Formulas: Test Reliability
IDENTIFIERS *Answer Until Correct: Distractors (Tests): *Test Instructions

ABSTRACT

A new scoring procedure for multiple choice tests attempts to assess partial knowledge and to restrict guessing. It is a variant of Coombs' elimination scoring method, adapted for use with the carbon-shield answer sheets commonly used with answer-until-correct scoring. Examinees are directed to erase the carbon shields of choices they are certain are incorrect (distractors). A scoring rule prohibits the examinee from uncovering the correct answer and then continuing to erase incorrect answers: once the correct choice is uncovered, no further erasures are made. This method was tried out on college-level art history students. Results indicated improved reliability and validity for the proposed method, as compared to number-right scoring and to Coombs' elimination scoring. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED190639

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A NEW METHOD FOR ADMINISTERING AND SCORING MULTIPLE-CHOICE TESTS: THEORETICAL CONSIDERATIONS AND EMPIRICAL RESULTS

Lawrence H. Cross

Nina J. Thayer

Robert B. Frary

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L C Ross

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Virginia Polytechnic Institute and State University

Blacksburg, Virginia

Paper presented at the meeting of the National Council on Measurement in Education, Boston, April, 1980

Session 16A

TM 800434

ABSTRACT

A new scoring procedure for multiple-choice tests is introduced which represents an attempt to assess partial information and restrict guessing. It is a variant of the elimination scoring method proposed by Coombs adapted for use with carbon shield answer sheets. Examinees are directed to erase the carbon shields of choices they are certain are incorrect. A scoring rule is imposed that will insure if the correct choice is uncovered, no further erasures will be made. Theoretical considerations are discussed and empirical results are presented which show substantial improvements in both reliability and validity associated with the proposed method over number-right and Coombs elimination scoring.

A NEW METHOD FOR ADMINISTERING AND SCORING MULTIPLE-CHOICE TESTS: THEORETICAL CONSIDERATIONS AND EMPIRICAL RESULTS

A large number of response/scoring methods have been proposed to replace number-right scoring of multiple-choice tests. Each of the proposed methods represents an attempt to improve the test score characteristics by minimizing the effect of guessing, assessing partial information, or both. Unfortunately, the empirical evidence to date does not generally support the advantages anticipated *a priori* by the proponents of each method. Most of the alternative procedures require an examinee to evaluate his level of assurance regarding the correctness of the choices offered and thereby run the risk of introducing a reliable, but invalid source of variance associated with personality factors. This is true of all confidence weighting procedures, the elimination scoring procedure proposed by Coombs (1953) as well as the conventional correction-for-guessing. This potential problem is avoided only by choice-weighted scoring and "answer-until-correct" (AUC) procedures wherein the response strategy that will optimize the examinee's expected score is straight-forward and does not require subjective decision making. However, neither of these methods is designed to minimize guessing, a limitation frequently cited by critics of multiple-choice tests. Of all of the alternative procedures, the method that obstensibly controls guessing most directly is the elimination score procedure wherein one point is awarded for each

distractor correctly identified and a penalty of (c - 1), where c is the number of choices, is imposed if the correct choice is inadvertently marked. As noted by Coombs, Milholland, and Womer (1956):

> Suppose, for example, an individual has crossed out two distractors he knows to be wrong on an item and has a sure two points to his credit. There remain two choices, one the answer and the other a distractor. If he has no knowledge and guesses, he is gambling an additional point credit versus a three point loss on a 50-50 chance.

While such an argument is convincing for an examinee possessing sufficient partial information to eliminate two out of four choices, the gain/loss ratio is reversed when an examinee is totally ignorant. If such an examinee decided to guess at random among all choices, he would be risking a maximum one-point loss against a possible three-point gain. While, in the long run, the expected score gain would be zero, luck may in the short run make such a strategy worthwhile in some situations. In any event, the deterent to extensive guessing in the face of ignorance is not great and amounts to no more than a correction, rather than a penalty for guessing. The greatest penalty is exacted when misinformation causes an examinee to mark only the correct choice as a distractor. In contrast when a person is quite confident that a distractor is the correct choice, and marks all other choices including the correct choice, the minimum penalty of a one pcint loss is imposed even though this state of knowledge may be the most troublesome. It was concerns of this type, which arose in working with the Coombs mode, that lead the first author to develop the new scoring method proposed in this paper.

Proposed Scoring Method

The new scoring method represents a variant of the elimination scoring method adapted for use with the carbon-shield answer sheets commonly used with AUC scoring. Specifically, examinees are directed to erase the carbon shields of only those choices they are certain are distractors. If the correct choice is uncovered, no further erasures are to be made. For five-choice items, a nine-point scoring system is used which assures compliance with instruction. One point is awarded for each choice left unerased regardless of the number or correctness of other erasures. Thus, five points are awarded for a question left unanswered. For each distractor correctly identified two points are awarded unless the correct choice is uncovered, in which case credit is awarded only for those choices left unerased. Thus a maximum score of nine is obtained by erasing four distractor shields (8 points) and leaving the correct choice shield unerased (1 point). The minimum score is obtained by erasing four shields including the shield of the correct choice thereby receiving one point for the last distractor left unerased. Consequently, the greatest penalty is associated with expressing high confidence and being misinformed whereas a minimum penalty is imposed (a score of 4) if misinformation is associated with little confidence as to the correct answer.

Theoretical Considerations

One of the major limitations of the Coombs scoring rule noted by Wang and Stanley (1970) is that the response strategy that will permit examinees to maximize their expected scores may become exceedingly complex if the correct answer is not known. The proposed scoring shares this same potential difficulty unless it can be assumed that for a given item, all choices which are not recognized as distractors will be equally attractive. Under this assumption, Frary (1980) has demonstrated analytically that, for items with three to six choices, examinees will maximize their expected score only if they eliminate at least one choice for each item regardless of knowledge, but that guessing after the first erasure cannot be expected to benefit them and may result in a severe penalty. To illustrate the correctness of this recommendation, consider a five choice item for which an examinee cannot with certainty eliminate even one choice. He should identify one distractor since a one point gain is being wagered against a one point loss when the probability of success is .80. Thus the expected gain would be .60, resulting in an expected item score of 5.6. If however, the examinee decides to guess at another choice his expected gain would be zero. If two distractors are known and a random guess were made among the remaining three choices, the expected gain would be -1 points.

Advising examinees to eliminate at least one choice for each item, even when they perceive such a response to be a random guess, is a desirable characteristic for two reasons. First, empirical evidence reported by Cross and Frary (1976) indicates that examinees are more successful than would be expected by chance when forced to guess on questions about which they claim to be totally ignorant. Consequently the proposed scoring system forces examinees to use this partial information that examinees seem to possess but do not recognize. Second, while the method encourages examinees to guess on the first erasure even when they perceive they have no information, the scoring rule strongly discourages such examinees from falsely expressing high confidence by marking out multiple distractors. Such a strategy cannot be expected to benefit an examinee after guessing on the first choice even though the correct answer may be revealed before the maximum penalty is imposed.

The same cannot be said about the Coombs scoring rule since on a five choice item an expected gain of zero points is associated with marking one choice or four choices at random, given the earlier stated assumption regarding the attractiveness of choices. Consequently there is no deterent to expressing high confidence in the face of ignorance in the hope that luck, in the short run, will be in the examinee's favor.

If, under either scoring rule, a correct choice is identified as a distractor, it is never certain whether it was the result of an

unsuccessful guess or misinformation. This uncertainty is especially of concern under the Coombs response made when the maximum number of choices is marked. Then the same minimum penalty is applied for both random guessing under ignorance and complete confidence that a distractor is the answer.

Under the proposed new mode, when an examinee has erased four shields, including the shield of the correct answer, it is clear that the last erasure uncovered the correct choice. Because such an error results in a loss of seven points on a five choice item, against a potential gain of only one point, misinformation is a far more reasonable explanation for such an outcome than guessing.

While the recommended directions discourage guessing beyond the first erasure per item, it may occur to a testwise individual that it may be advantageous to ignore the directions and guess before erasing the shields of those choices he is certain are incorrect, thereby minimizing the loss. For example, suppose an examinee has assigned subjective probabilities of correctness of .6 and .4 to two out of the five choices. It might appear advantageous to erase the shield of the latter choice before claiming credit for the three choices for which a zero personal probability of correctness is assigned. However, preliminary work in this area suggests that such a strategy will not be advantageous unless the personal probabilities are extremely disparate across the choices of which the examinee is uncertain. For example, if probabilities of .8 and .2 have been

assigned for two choices of a five choice item or .9 and .1 for a four choice item the above strategy may be advantageous. However, it seems unlikely that examinees can make such fine discriminations, and it seems reasonable that most examinees in this situation would not consider the elimination of a choice associated with a very low probability to represent a guess.

The proposed method also differs considerably from the Coombs method in the way expressions of misplaced confidence are penalized. If the maximum number of choices have been identified as distractors, including the correct choice, it is most likely to represent complete misinformation under the proposed scoring rule. In a "real world" situation, this state of knowledge is likely to be more troublesome than partial misinformation wherein the examinee is still unsure of the correctness of the remaining choices in spite of having classified the answer as incorrect. Whereas complete misinformation is likely to lead to inappropriate action, a person possessing partial misinformation and yet lacking confidence as to the correct answer would presumably seek further information before taking action. Given this rationale, it seems appropriate to impose the maximum penalty for complete misinformation and the minimum penalty for partial misinformation, as is the case with the proposed scoring rule. In contrast, the Coombs method reverses the magnitude of penalties by imposing the least penalty for an expression of complete misinformation, which just

as well may represent a bold guess in the face of ignorance. Consequently, the proposed scoring system not only appears to provide a more effective control over bold guessing and can more effectively distinguish a bold guess from complete misinformation but also penalizes misinformation in what appears to be a more appropriate manner than does the Coombs method.

In addition to treating misinformation and guessing in a manner that appears more reasonable than that provided by the Coombs mode, the proposed scoring method shares with AUC scoring the advantage of providing feedback at the time of testing and thereby affords examinees the opportunity to determine their scores before turning in their answer sheets.

A Preliminary Investigation

Method

Empirical evidence regarding the reliability and validity of scores resulting from the proposed method of scoring was obtained by the second author as part of a larger study. Of primary interest to this report is a comparison of the score characteristics resulting from responses to the same multiple-choice tests by two intact groups of college students, one group trained in the use of the Coombs method and the other trained in the use of the proposed method. These two groups constituted classes of an art history course at a medium-sized southwestern Virginia university. A series of three, five-choice

multiple choice tests were administered to each group during the winter 1979 semester. Both classes responded to the first quiz in the conventional manner with number-right scoring, but for the second and third tests, the alternative response modes were used. Each class was provided detailed oral and written instructions in the use of one or the other method. Sample items and examples of how selected response patterns for an item would be scored were illustrated. However, students were not advised as to the response strategy they should use to maximize their expected scores. The -4 to 4 scoring rule which was explained to the students using the Coombs mode, was subsequently transformed to a 1 to 9 scale to facilitate comparison with the 1 to 9 scale used with the proposed method.

Results

The descriptive statistics, as well as relability and validity estimates for the two groups, are presented in Table 1. Reliability was estimated by coefficient alpha and validity was extimated using

[Insert Table 1 here]

college grade-point average as the criterion. The difference in mean number-right scores on the first quiz suggests that group 2 (proposed method) was of higher ability, an observation supported by a difference in the same direction associated with mean grade-point averages (2.60 vs 2.44). Consequently, the test score characteristics

must be qualified as a result of using intact, rather than randomly formed groups. However, this apparent difference in group ability would tend to favor enhanced reliability and validity for scores earned under the Coombs mode (group 1) since the average item difficulty for this group more nearly approximated the optimal value of .5. Despite this limitation, the results reveal elevated reliability and validity coefficients associated with scores obtained by the proposed scoring method over Coombs scores on both tests.

While the coefficients reported do not represent statistically significant differences (the w statistic proposed by Feldt (1969) was used to test alpha coefficients) as might be expected for such small sample sizes, it is of interest to evaluate the coefficients descriptively by considering changes in test length required to account for such differences. With respect to reliability, use of the Spearman/Brown prophecy formula suggests that the second and third quizzes would have to be increased in length by a factor 1.5 and 1.3 respectively if the Coombs mode scoring was to achieve the level of reliability obtained under the proposed scoring method. With respect to validity, use of the Spearman/Brown formula suggests that increases in test length alone could not be expected to raise the validity coefficient of the Coombs mode scores to the levels associated with the proposed scoring method. Alternatively, using the reliability and validity coefficients associated with the proposed scoring method as the basis for comparison, it is estimated that quizzes two

and three could be reduced in length by 30 and 21 percent respectively before the validity coefficients would dropoff to the level associated with the Coombs method. Of course all of the above comparisons may reflect intact group differences other than these related to the scoring methods.

It should also be noted that although both methods of scoring are based upon a nine-point scale, the standard deviation of the scores are much larger for the proposed method than for the Coombs method on both tests. Part of this difference no doubt is due to the different manner in which guessing and misinformation are treated by the two methods, and apparently a disproportionate amount of this increase in variance is true score variance.

While a direct comparison of the alternative scoring methods to number-right scoring is not possible since different tests are involved, it is of interest to note that the tests scored with the alternative methods were far superior to the test scored with number-right scoring in predicting the criterion; an outcome which cannot be attributed to differences in reliability.

Conclusion

The proposed scoring system represents an attempt to assess partial information and at the same time restrict guessing without unduly penalizing examinees possessing partial misinformation. A desirable feature of the scoring rule is that examinees should be

directed to identify at least one choice as a distractor for each item but that beyond that, any further guessing cannot be expected to benefit the student and may result in a severe penalty. To the extent that the method is successful in minimizing guessing and assigning appropriate score credit to various levels of information and misinformation, it should enhance test score characteristics over number-right and Coombs mode scoring. Preliminary empirical data suggests that such may be the case, but, because of design deficiencies, the outcomes reported must be viewed as indicative rather than definitive of the potential of the proposed scoring method to achieve these goals. Further research is essential before recommendations for practical use can be made.

References

Coombs, C. H. On the use of objective examinations. Educational and Psychological Measurements, 1953, 13, 308-310.

Coombs, C. H., Milholland, J. E., and Wormer, F. B. The assessment of partial knowledge. Educational and Psychological Measurement, 1956, 16, 13-37.

Cross, L. H. and Frary, R. B. An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. Journal of Educational Measurement, 1977, 14, 313-321.

Feldt, L. S. A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. Psychometrika, 1969, 34, 363-373.

Frary, R. B. The effect of misinformation, partial information and guessing on expected multiple-choice test item scores. Applied Psychological Measurement. In press, 1980.

Wang, M. W. and Stanley, J. C. Differential weighting: A review of methods and empirical studies. Review of Educational Research, 1970, 40, 663-705.

Table 1

Descriptive Test Statistics for Two Groups Taking the Same Three Tests Under Different Scoring Methods

	Quiz #	# of Items	Scoring Method	Mean	S.D.	$r_{xx}(\alpha)$	$r_{xy}(y = GPA)$
Group 1 n = 37	1	25	NR	16.05	3.50	.648	.309
	2	18	Coombs	119.35	13.55	.601	.524
	3	20	Coombs	131.31	13.83	.504	.521
Group 2 n = 37	1	25	NR	19.08	3.02	.651	.136
	2	18	Proposed	127.16	19.12	.693	.684
	3	20	Proposed	145.37	16.79	.585	.706